

Minimax quantum tomography: the ultimate bounds on accuracy

Christopher Ferrie^{1,2} and Robin Blume-Kohout³

¹Center for Quantum Information and Control, University of New Mexico, Albuquerque, New Mexico, 87131-0001

²Centre for Engineered Quantum Systems, School of Physics,
The University of Sydney, Sydney, NSW, Australia

³Sandia National Laboratories, Albuquerque, New Mexico, 87185

(Dated: March 12, 2015)

A *minimax* estimator has the minimum possible error (“risk”) in the worst case. We construct the first minimax estimators for quantum state tomography with relative entropy risk. The minimax risk of non-adaptive tomography scales as $O(1/\sqrt{N})$, in contrast to that of classical probability estimation which is $O(1/N)$. We trace this deficiency to *sampling mismatch*: future observations that determine risk may come from a different sample space than the past data that determine the estimate. This makes minimax estimators very biased, and we propose a computationally tractable alternative with similar behavior in the worst case, but superior accuracy on most states.

Quantum information processing relies on physical *qubits* that store and process quantum information. Testing and characterizing qubit devices is the business of quantum tomography [1], and *quantum state tomography* in particular is used to estimate the quantum state (density matrix) ρ produced by an initialization procedure. Tomography comprises two steps: (1) *data gathering*, accomplished by measuring a “quorum” of different observables on N samples of ρ ; and (2) an *estimator* that maps the data to a final estimate $\hat{\rho}$. The goal, of course, is an accurate estimate – we want $\hat{\rho}$ to be “close” to the true state ρ , minimizing some error metric $d(\rho : \hat{\rho})$.

One might thus expect that tomographers would choose an estimator that is optimal (or at least near-optimal) in accuracy. Somewhat surprisingly, this is not done. Although several estimators are known and used (linear inversion [2], maximum likelihood [3], Bayesian mean [4], hedged maximum likelihood [5], L_1 -regularization [6]), none of them is known to have optimal pointwise accuracy [22] for finite N . In fact, we don’t even know the ultimate bounds on accuracy, which makes it impossible to say which of these estimators (if any) are “good enough”.

We remedy this embarrassing situation in the present Letter by constructing *minimax* estimators (depicted in Fig. 1; see detailed explanation after Eq. 7) with absolutely optimal performance. These estimators are unwieldy, but (i) their performance yields tight upper bounds on accuracy, effectively delineating what “good enough” means, and (ii) their construction provides quite a lot of insight into the structure of the problem. Armed with these results, we show that *hedged maximum likelihood* (HML) is remarkably close to optimal, and outperforms minimax for most states (though of course its worst-case risk is higher). We also identify a good value for the hedging parameter β that appears in HML.

Prerequisites: Defining “optimal” requires making several choices. For example, an optimal estimator for one error metric $d(\rho : \hat{\rho})$ is generally not optimal for a different metric $d'(\rho : \hat{\rho})$. Here [7], we quantify inaccuracy by

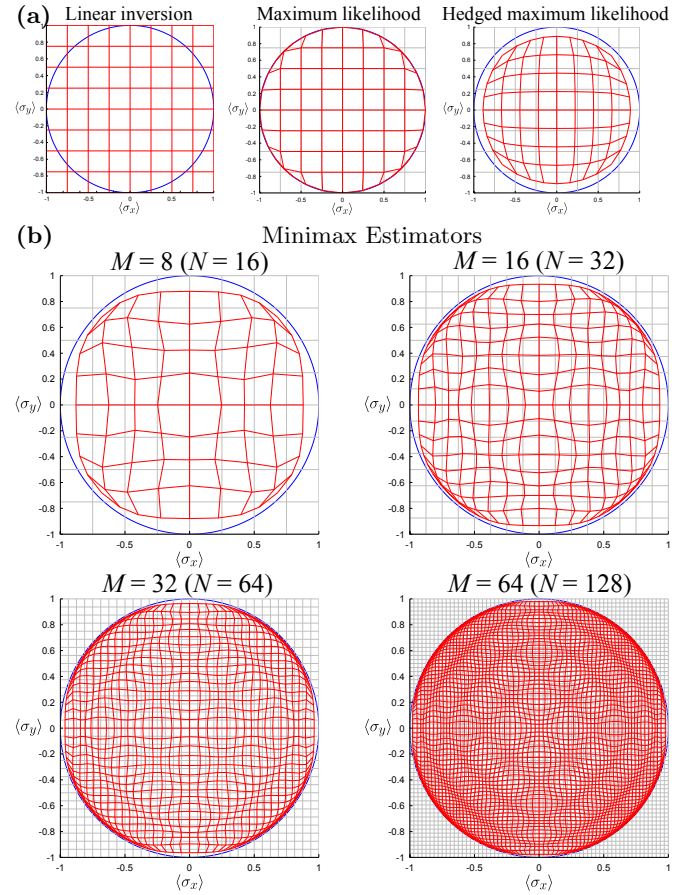


FIG. 1: **Estimators for Pauli measurements on a rebit**, depicted as distortions of the “linear inversion grid” (see text after Eq. 7). (a) Three standard estimators, each for $M = 8$ measurements of X and Y . Each vertex of the red grid corresponds to an estimated density matrix. Linear inversion estimates may lie outside “Bloch disk” of physical states. MLE estimates are non-negative, while HML yields strictly positive estimates. (b) Minimax estimators for $M = 8, 16, 32, 64$ measurements of X and Y on a rebit. They are locally biased, toward support points of the least favorable prior.

the *quantum relative entropy*,

$$d(\rho : \hat{\rho}) = D(\rho || \hat{\rho}) = \text{Tr} [\rho(\log \rho - \log \hat{\rho})]. \quad (1)$$

Like its classical analogue, quantum relative entropy [21] is a well-motivated measure of *predictive* (and information-theoretic) inaccuracy [4]. It quantifies the expected cost, resulting from an imperfect estimate, of imperfectly predicting measurements of ρ 's diagonal basis (because this is the hardest measurement to predict accurately).

An estimator's *pointwise risk* is a function of the true state ρ and is given by the average of $d(\rho : \hat{\rho})$ over all possible data sets D :

$$\bar{d}(\rho) = \sum_D \text{Pr}(D|\rho) d(\rho : \hat{\rho}(D)). \quad (2)$$

In the minimax paradigm, we quantify an estimator's accuracy by its *worst-case risk*, $\bar{d}_{\max} = \max_{\rho} \bar{d}(\rho)$. The *minimax risk* of the estimation problem is the minimum achievable risk (minimized over all possible estimators), and a *minimax estimator* is one that achieves this bound.

In most inference problems, the sample space of possible observations (data) is fixed by the problem. Not so in quantum tomography. Quantum systems can be measured in many different and incomparable ways. This is the single most significant difference between quantum and classical estimation. This freedom is often removed in quantum problems by choosing the best or worst possible measurement (e.g., as in the definition of quantum relative entropy as the classical relative entropy of the most difficult-to-predict measurement). This is usually not done in tomography, because the optimal measurements are far too difficult. In this letter, we follow the majority of experiments and analyze tomography based on Pauli measurements on a single qubit. However, we also prove analytic lower bounds on minimax risk that apply to *any* non-adaptive measurement and any d -dimensional quantum system. In some parts of our analysis, we use a *rebit* – a quantum system with a 2-dimensional *real* Hilbert space, whose state space corresponds to the equatorial plane of the Bloch sphere – as an easier-to-analyze proxy for a qubit.

Minimax risk: The first main result of this Letter is a lower bound on the asymptotic ($N \rightarrow \infty$) minimax relative entropy risk of Pauli tomography on qubits and rebits,

$$\bar{d}_{\max} \geq \frac{e^{-\frac{1}{2}} \sqrt{D-1}}{4 \sqrt{N}}, \quad (3)$$

where $D = 2$ for rebits and $D = 3$ for qubits. Its $O(1/\sqrt{N})$ scaling contrasts sharply with the minimax risk of estimating a *classical* bit, which is almost exactly $0.5/N$ [10, 11]. We derive this bound below by mapping the minimax risk of qubit and rebit state tomography to

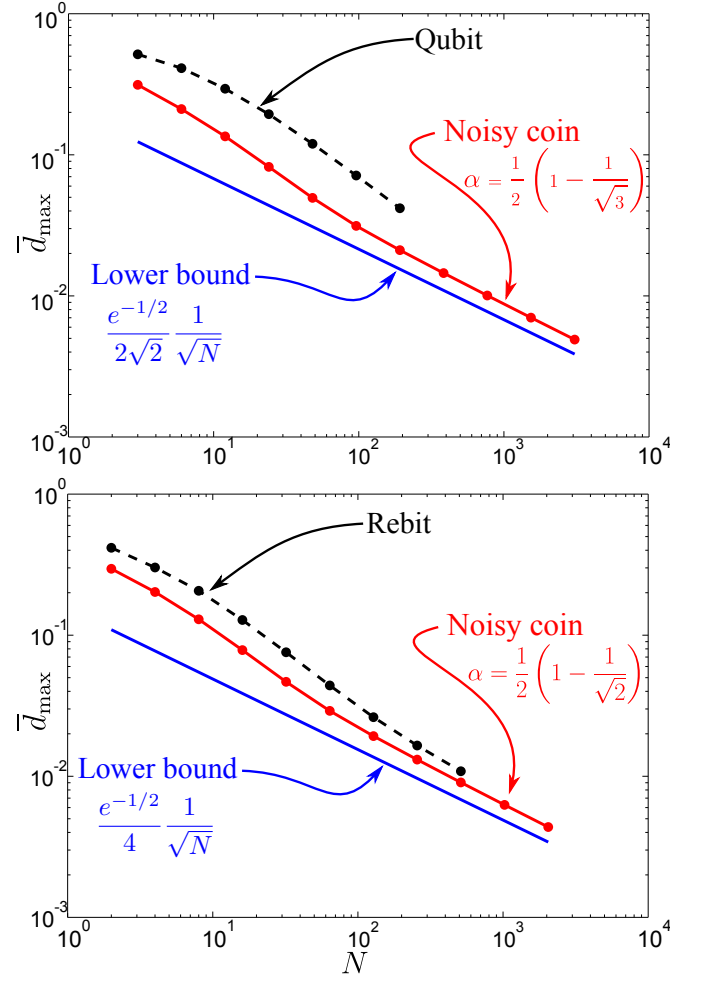


FIG. 2: **Numerical minimax risk for qubits, rebits, and noisy coins.** Black curves show the risk of numerically constructed minimax estimators for (a) a qubit and (b) a rebit, as a function of the number of samples (N), up to the maximum that was numerically feasible. Red curves illustrate the numerically-computed risk of “noisy coin” systems whose noise levels are chosen to match the effective “noise” of the qubit and the rebit (respectively). Blue lines show the lower bound given in Eq. (6).

a classical “noisy coin” model. In Figure 2, we compare these bounds to numerical calculations of the minimax risk, for small N , of qubits, rebits, and noisy coins.

A d -dimensional quantum state is analogous in many ways to a classical d -outcome probability distribution. However, its minimax risk scales differently because of a phenomenon intrinsic to quantum tomography (though not uniquely quantum) that we call *sampling mismatch*: the sample space for the observed events is neither unique nor isomorphic to the underlying state space. For example, the possible statistics for the three 2-outcome Pauli measurements on a qubit naturally define a cube, whereas the possible *quantum* states form a sphere (the Bloch ball).

Sampling mismatch can be reproduced in a simple classical model called the “noisy coin” [12]. It is a classical system with a 2-outcome sample space (i.e., a coin flip) where each observation is erroneous with known probability α . Sampling mismatch arises when we attempt to assign probabilities to future *noiseless* observations using data from *noisy* measurements. The noisy coin’s minimax risk is $O(1/\sqrt{N})$, because nearly-pure states are hard to estimate accurately from noisy statistics. The corresponding minimax estimators are strongly biased toward nearly-pure states (see [12] for details). We are going to use a variant of the noisy coin model to bound the risk of tomography.

We define “tomography” thus: N samples (copies) of a single-qubit state ρ will be prepared; each sample will be measured independently (not jointly together with other samples) in a predefined fashion (not adaptively). The k th sample is measured in an arbitrary basis, and this measurement can be described by a POVM (positive operator-valued measure) $\mathcal{M}_k = \{\Pi_k, \mathbb{I} - \Pi_k\}$ whose outcomes have probabilities $\{q, 1 - q\}$ with $q = \text{Tr} \Pi_k \rho$. Based on the N measurement results, we report a state $\hat{\rho}$, and seek to minimize relative entropy cost.

Now, suppose that before analyzing the data (but after choosing the measurements!) we are told the eigenbasis of ρ . This helps us (only ρ ’s spectrum must be estimated), so the risk of spectrum estimation is a strict *lower* bound on the risk of full tomography[23].

We define $\{|0\rangle, |1\rangle\}$ to be the eigenstates of ρ , and write

$$\rho = p |0\rangle\langle 0| + (1 - p) |1\rangle\langle 1|. \quad (4)$$

Now, we need only estimate $p \in [0, 1]$. This parameter manifold is identical to that of a coin. Furthermore, the quantum relative entropy between two diagonal density matrices is identical to the classical relative entropy between the corresponding distributions. So, since ρ ’s eigenbasis is known, estimating ρ is identical to estimating the bias of a coin. However, unless the eigenbases of ρ and the Π_k happen to coincide, the measurement data obtained from the N samples of ρ are not “noiseless”. Even if $p = 0$ (i.e., ρ is pure), the data remain somewhat random. The probability of observing Π_k is not p , but

$$\begin{aligned} q &= p \langle 0 | \Pi_k | 0 \rangle + (1 - p) \langle 1 | \Pi_k | 1 \rangle \\ &= p(1 - 2\alpha_k) + \alpha_k \end{aligned}$$

where the *effective noise* in sample k is

$$\alpha_k = \langle 1 | \Pi_k | 1 \rangle^2. \quad (5)$$

We can model this situation perfectly by a noisy coin (as in Ref. [12]) where each observation fails with a different error probability. The error probability for the k th sample is α_k . In the appendix, we bound this estimation

problem’s minimax risk by

$$\bar{d}_{\max} \geq \frac{e^{-\frac{1}{2}}}{2\sqrt{\bar{\beta}}} \frac{1}{\sqrt{N}}, \quad (6)$$

where $\bar{\beta}$ is the average *resolution* provided by the N noisy samples:

$$\bar{\beta} = \frac{1}{N} \sum_{k=1}^N \beta_k = \frac{1}{N} \sum_{k=1}^N \frac{(1 - 2\alpha_k)^2}{\alpha_k(1 - \alpha_k)}. \quad (7)$$

For any fixed measurement strategy – e.g., the standard one where $N/3$ samples are measured in the X, Y, Z bases – the *maximum* risk occurs when we choose the eigenbasis of ρ to maximize $\bar{\beta}$ in Eq. 7. This “least favorable” basis is the one that lies as far as possible from all measured bases. For a rebit, it lies halfway between the X and Z bases, and $\alpha_k = \frac{1}{2}(1 - 1/\sqrt{2})$. For a qubit, it is the geometric mean of the X, Y , and Z bases, and $\alpha_k = \frac{1}{2}(1 - 1/\sqrt{3})$. Inserting these values for α_k yields the final bound given in Eq. 3.

This argument applies (qualitatively) to tomography on any finite-dimensional system with any discrete POVM. As long as no samples are measured in a basis that diagonalizes ρ , the minimax risk scales as $O(1/\sqrt{N})$ (although the prefactor will vary). However, if any non-vanishing fraction of the N samples are measured in a basis that diagonalizes ρ , then Eq. 6 no longer applies. Thus, continuous POVMs such as the unitarily invariant Haar-uniform rank-1 POVM (a.k.a. the uniform POVM), require a slightly different argument. In the appendix, we prove that even in this case, the minimax risk is lower bounded by $O((N \log N)^{-1/2})$.

Estimators: To confirm the bound given by Eq. 3 and explore minimax risk at small N , we use numerics to find minimax estimators. An estimator is a map from the set of all possible datasets into the set of density matrices. The outcomes of the measurement(s) performed are represented by a set of positive operators $\{E_k\}$, and the data themselves by a set of frequencies $D = \{n_k\}$. For qubit Pauli tomography, the data comprise $M = N/3$ samples each of σ_x, σ_y , and σ_z measurements; for rebits, they comprise $M = N/2$ samples each of σ_x and σ_y measurements.

We used numerical optimization (over the set of possible estimators) to find minimax estimators. The algorithms are described in the appendix. In Figure 1, we depict the resulting estimators, and compare them to three canonical estimators:

1. **Linear inversion** ($\hat{\rho}_{\text{LI}}$): The first tomographic estimator, it is obtained by equating each probability $\text{Pr}(k|\hat{\rho}_{\text{LI}}) = \text{Tr} E_k \hat{\rho}_{\text{LI}}$ to its observed frequency $\frac{n_k}{M}$.
2. **Maximum likelihood** ($\hat{\rho}_{\text{ML}}$): MLE assigns the density matrix that maximizes the probability of the observed data (the likelihood), $\mathcal{L}(\rho) = \text{Pr}(D|\rho) = \prod_k [\text{Tr}(E_k \rho)^{n_k}]$.

3. Hedged maximum likelihood ($\hat{\rho}_{\text{HML},\beta}$): The HML estimator maximizes the product of $\mathcal{L}(\rho)$ and a “hedging function” $h(\rho) = \det(\rho)^\beta$. This function is strictly convex and vanishes for rank-deficient states, so the HML estimate is always full-rank.

To simplify visualization, we depict *rebit* estimators, which are qualitatively similar to qubit estimators and easier to depict. A rebit estimator is a map from datasets to Bloch vectors, as $\hat{\rho} : \{0, \dots, M\}^2 \rightarrow \mathbb{R}^2$. We use the linear inversion estimator as a reference. As a linear map from the 2-dimensional space of datasets ($\{0 \dots M\}^2$) and the 2-dimensional space of rebit states (the unit disc in \mathbb{R}^2), the linear inversion estimator is represented by a uniform grid on the “Bloch square” (Fig. 1a). Every *other* estimator is represented as a distortion of this grid. The vertices of the grid are estimates $\hat{\rho}$, and the position of such a vertex within the grid indicates what dataset it came from.

Minimax estimators for $N = 16, 32, 64$ and 128 (total) Pauli measurements on a rebit are shown in Figure 1b. The most striking feature of these estimators is a pronounced “ripple” phenomenon. This is not a numerical artifact. Instead, it represents a consistent bias toward certain discrete points within the state space (support points of the least favorable prior – see Fig. 4 in the appendix), which can be identified in Figure 1 as regions where the grid lines cluster together. The minimax estimator demonstrates this bias because these points are, in a particular sense, the most difficult to estimate accurately.

Improving on Minimax: The minimax criterion is an elegant concept, but a dangerous one. In its single-minded quest to improve the *maximum* risk, it has no concern for the pointwise risk at states that are “easier” to estimate. In such regions, it may incur extreme bias and inaccuracy, for the sole purpose of achieving a tiny reduction in the maximum risk. For quantum tomography, this effect become extreme. While $O(1/N)$ risk can be achieved on all full-rank states, the risk is unavoidably $O(1/\sqrt{N})$ near the boundary. Our numerical experiments confirm that the minimax estimator’s pointwise risk is $O(1/\sqrt{N})$ everywhere, whereas other estimators easily achieve $O(1/N)$ risk in the interior of the Bloch sphere (Fig. 3b). If ρ really was selected adversarially, then minimax would be a wise strategy. But in realistic cases, we would prefer an estimator that achieved $O(1/N)$ scaling where possible, even at the cost of slightly worse *worst-case* behavior.

A good estimator should achieve $O(1/N)$ risk in the interior, while coming as close as possible to minimax performance near the boundary. The maximum likelihood estimator (MLE) is disqualified because its pointwise expected risk is uniformly infinite (it has nonzero probability of returning a rank-deficient estimate for every ρ , so $\bar{d}(\rho) = \infty$). However, *hedged maximum likeli-*

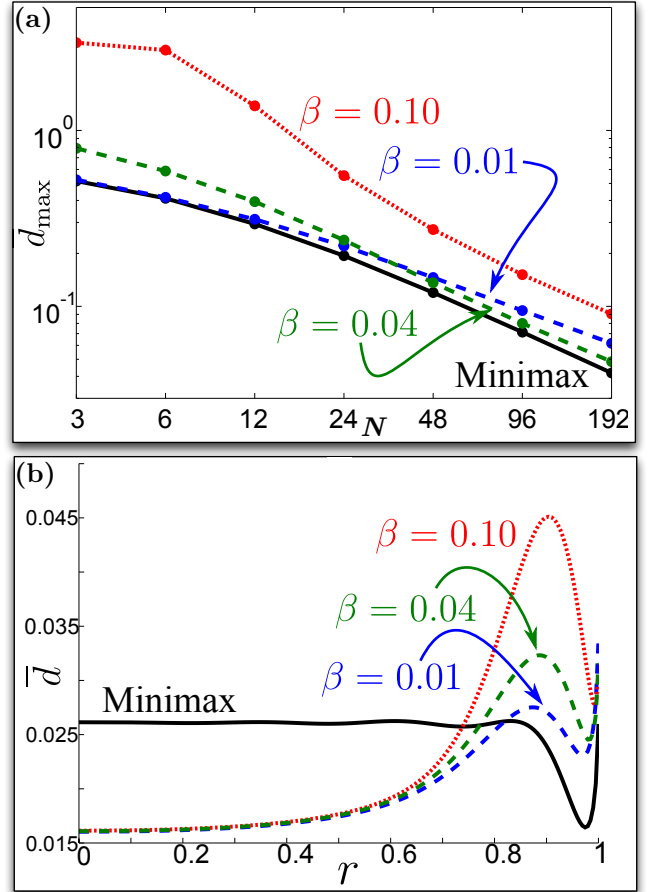


FIG. 3: **Maximum and pointwise risk of minimax and HML estimators.** Plot (a) shows the maximum risk, for *qubit* tomography, of the minimax estimator and three different HML estimators ($\beta = 0.01, 0.04, 0.10$) for $N \leq 192$ samples distributed equally among the 3 Pauli bases. Plot (b) shows the pointwise risk, along the axis oriented at 45 degrees to both X and Y , of the same estimators for $N = 128$ samples for a *rebit* (this minimax estimator is depicted in Fig. 1b). The two local maxima of $\bar{d}(\rho)$ are at $r = 1$ and $r \approx 1 - \frac{1}{\sqrt{N}}$. Choosing $\beta \approx 0.04$ balances these risks, and is therefore minimax among HML estimators. This optimal HML estimator comes very close to matching the worst-case performance of the minimax estimator, and outperforms it dramatically in the interior of the state space.

hood (HML) does not have this behavior. Introduced in Ref. [5] as a full-rank alternative to MLE, HML generalizes classical “add- β ” estimators. Like them, it never assigns zero probabilities, and has an adjustable parameter β that governs how much it avoids zero eigenvalues. Classical “add- β ” estimators are very nearly minimax (for $\beta \approx 1/2$), which suggests that HML estimators might have similar near-optimality properties.

All HML estimators have good behavior ($O(1/N)$ pointwise risk) in the interior, so we are free to define the “optimal” β by minimax (among HML estimators). As illustrated in Fig. 3b, an HML estimator’s pointwise risk

has local maxima at the boundary (pure states) and/or at a slightly depolarized state (with purity $\sim 1 - 1/\sqrt{N}$). To minimize its maximum, we choose β to equalize the risk at these two local maxima. The asymptotically optimal β for the noisy coin model was shown in Ref. [12] to be $\beta_{\text{optimal}} \approx 0.0389$, and our numerics confirm that $\beta \approx 0.04$ is optimal to within the available numerical precision for rebit tomography as well (Fig. 3b; qubit results for smaller N are not shown, but confirm that $\beta \approx 0.04$ has nearly-minimax performance).

For this near-optimal value of β , HML compares favorably with minimax estimators. Its worst-case risk is very close to the minimax risk (Fig. 3a), and it dramatically outperforms minimax in the interior of the state space (Fig. 3b). So while optimal hedging estimators do not offer strictly optimal performance by any criterion, they are (i) easy to specify and calculate, (ii) close to minimax, and (iii) more accurate than minimax estimators for almost all states ρ . We do not know why the optimal β is so different for noiseless coins (≈ 0.5) and for qubits/rebits/noisy coins (≈ 0.04), but it suggests fundamental differences between noiselessly sampled systems and those (like qubits and noisy coins) where sampling mismatch is important.

CF was supported by National Science Foundation grant number PHY-1212445, the Canadian Government through the NSERC PDF program, the IARPA MQCO program, the ARC via EQuS project number CE11001013, and by the US Army Research Office grant numbers W911NF-14-1-0098 and W911NF-14-1-0103. Sandia National Laboratories is a multi-program laboratory operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

[1] M. Paris and J. Rehacek, *Quantum state estimation*. Springer, New York (2004).
[2] M. A. Nielsen and I. L. Chuang, *Quantum computation and quantum information*. Cambridge University Press (2010).
[3] Z. Hradil, *Quantum-state estimation*, *Physical Review A* **55**, R1561 (1997).
[4] R. Blume-Kohout, *Optimal, reliable estimation of quantum states*, *New Journal of Physics* **12**, 043034 (2010).
[5] R. Blume-Kohout, *Hedged maximum likelihood quantum state estimation*, *Physical Review Letters* **105**, 200504 (2010).
[6] D. Gross, Y.-K. Liu, S. T. Flammia, S. Becker, and J. Eisert, *Quantum state tomography via compressed sensing*,

Physical Review Letters **105**, 150401 (2010).
[7] Without a doubt there are other reasonable definitions that would lead to different answers. Examples are given by Refs. [16, 17], who each obtain substantially different conclusions by choosing different loss functions. We expect future work to consider other variations, and we hope that our results and discussion will help to inform and guide such future work.
[8] E. L. Lehmann and G. Casella, *Theory of point estimation*, Springer (1998).
[9] A rebit is a system with a 2-dimensional *real* Hilbert space, containing states $|\psi\rangle = \sin\theta|0\rangle + \cos\theta|1\rangle$. Alternatively, consider a qubit with the constraint $\langle\sigma_y\rangle = 0$.
[10] A. Rukhin, *Minimax estimation of the binomial parameter under entropy loss*, *Statistics and Decisions* **13**, 69 (1993).
[11] R. E. Krichevskiy, *Laplace's law of succession and universal encoding*, *IEEE Transactions on Information Theory* **44**, 296 (1998).
[12] C. Ferrie and R. Blume-Kohout, *Estimating the bias of a noisy coin*, *AIP Conference Proceedings* **1443**, 14 (2012).
[13] D. Mahler, L. A. Rozema, A. Darabi, C. Ferrie, R. Blume-Kohout, and A. Steinberg, *Adaptive quantum state tomography improves accuracy quadratically*, *Physical Review Letters* **111**, 183601 (2013).
[14] E. Bagan, M. Ballester, R. Gill, R. Muñoz-Tapia, and O. Romero-Isart, *Separable measurement estimation of density matrices and its fidelity gap with collective protocols*, *Physical Review Letters* **97**, 130501 (2005).
[15] O. Barndorff-Nielsen and R. Gill, *Fisher information in quantum statistics*, *Journal of Physics A: Mathematical and General* **33**, 4481 (2000).
[16] H. K. Ng and B.-G. Englert, *A simple minimax estimator for quantum states*, *International Journal of Quantum Information* **10** (2012).
[17] S. T. Flammia, D. Gross, Y.-K. Liu and Jens Eisert, *Quantum tomography via compressed sensing: error bounds, sample complexity and efficient estimators*, *New Journal of Physics* **14** 095022 (2012).
[18] M. V. Burnashev and S.-i. Amari, *On density estimation under relative entropy loss criterion*, *Problems of Information Transmission* **38**, 323 (2002).
[19] P. J. Kempthorne, *Numerical specification of discrete least favorable prior distributions*, *SIAM Journal on Scientific and Statistical Computing* **8**, 171 (1987).
[20] M. Guță and J. Kahn, *Local asymptotic normality for qubit states*, *Physical Review A* **73**, 052108 (2006).
[21] V. Vedral, *The role of relative entropy in quantum information theory*, *Reviews of Modern Physics* **74**, 197 (2002).
[22] Bayesian mean is known to provide optimal accuracy on average over a known prior distribution of ρ , for certain important error metrics $d(\rho : \hat{\rho})$.
[23] Technically, the risk can be written as a sum of terms due to (i) spectral error, and (ii) eigenbasis error. It turns out that the minimax risk is always dominated by spectral error. Thus, giving up the eigenbasis error does not lower the risk substantially.

The minimax risk of a noisy coin

In this appendix we show that the minimax risk of estimating the bias of a *noisy* coin is $O(1/\sqrt{N})$ (in contrast to the $O(1/N)$ minimax risk for a noiselessly observed coin), and derive a simple lower bound on it.

Now, suppose a coin with bias $p = \Pr(\text{“heads”})$ is flipped N times and a sequence of binary outcomes $\mathbf{n} = \{n_k\}$ are recorded. But these observations are unreliable; each outcome is recorded incorrectly with trial-dependent probabilities $\alpha = \{\alpha_k\}$ (all taken from the interval $[0, \frac{1}{2})$). The distribution of the outcomes is

$$\Pr(\mathbf{n}|p, \alpha) = \prod_{k=1}^N \Pr(n_k|p, \alpha_k) = \prod_{k=1}^N q_k^{n_k} (1 - q_k)^{1-n_k}, \quad (8)$$

where the probability of observing “heads” on trial k is not p , but

$$q_k(p) = \alpha_k + p(1 - 2\alpha_k) = p + \alpha_k(1 - 2p). \quad (9)$$

We recover a standard noiseless coin when $\alpha_k = 0$ for all k .

For each prior distribution $\mu(p)$, the estimator with the smallest risk (expected cost) is the *Bayes estimator* for $\mu(p)$, and its risk is the *Bayes risk* of $\mu(p)$. Bayes estimators need not be simple, but because relative entropy is a *Bregman divergence*, the Bayes estimator is always the mean of the posterior distribution (obtained via Bayes’ Rule) [4]. The prior with the highest Bayes risk is the *least favorable* prior, and its risk is the minimax risk. Thus, the Bayes risk of any prior is a lower bound for the minimax risk, which suggests an obvious variational approach to bounding the minimax risk by choosing a prior whose risk is high but easy to calculate. Obviously, some priors have very low risk [e.g., $\mu(p) = \delta(p - p_0)$], and provide useless lower bounds. A common approach is to use the uniform (Lebesgue) prior, but for the noisy coin this prior actually has rather low $[O(1/N)]$ risk. So instead, we consider the set of *bimodal priors*,

$$\pi(p) = \frac{\delta(p - p_0) + \delta(p - p_1)}{2}. \quad (10)$$

(Varying the weights yields a slightly less favorable prior, but doesn’t change the asymptotic scaling). We will choose $p_0 = 0$ and $p_1 \approx 1/\sqrt{N}$.

The risk of π is given by $\bar{d}(\pi) = [\bar{d}(0) + \bar{d}(p_1)]/2$, and by observing that $\bar{d}(p_1) \geq 0$ we obtain the lower bound

$$\bar{d}(\pi) \geq \frac{1}{2}\bar{d}(0) = \frac{1}{2}\mathbb{E}_{\mathbf{n}|p=0}[D(0||\hat{p}(\mathbf{n}))], \quad (11)$$

where the Bayes estimator is the posterior mean, given by

$$\hat{p}(\mathbf{n}) = \frac{p_1 \Pr(\mathbf{n}|p_1)}{\Pr(\mathbf{n}|p_1) + \Pr(\mathbf{n}|0)} = \frac{p_1}{1 + \Lambda(\mathbf{n})}, \quad (12)$$

in terms of the *likelihood ratio*

$$\Lambda(\mathbf{n}) = \frac{\Pr(\mathbf{n}|0)}{\Pr(\mathbf{n}|p_1)}. \quad (13)$$

We can lower-bound the relative entropy term by $D(0||\hat{p}) = -\log(1 - \hat{p}) \geq \hat{p}$, so

$$\bar{d}(0) \geq \mathbb{E}_{\mathbf{n}|p=0}[\hat{p}(\mathbf{n})] = p_1 \mathbb{E}_{\mathbf{n}|p=0} \left[\frac{1}{1 + \Lambda(\mathbf{n})} \right]. \quad (14)$$

If we define $\lambda(\mathbf{n}) = -2 \log \Lambda(\mathbf{n})$ and apply Jensen’s inequality, we obtain

$$\bar{d}(0) \geq p_1 e^{-\frac{1}{2}\mathbb{E}_{\mathbf{n}|p=0}[\lambda(\mathbf{n})]}. \quad (15)$$

Next, we perform a Taylor expansion of the expectation $E_{\mathbf{n}|p=0}[\lambda(\mathbf{n})]$ around $p_1 = 0$. The derivatives of the likelihood function (8) are

$$\frac{\partial}{\partial p_1} \log \Pr(\mathbf{n}|p_1) = \sum_k \left(\frac{n_k(1 - 2\alpha_k)}{q_k} - \frac{(1 - n_k)(1 - 2\alpha_k)}{1 - q_k} \right), \quad (16)$$

$$\frac{\partial^2}{\partial p_1^2} \log \Pr(\mathbf{n}|p_1) = \sum_k \left(-\frac{n_k(1 - 2\alpha_k)^2}{q_k^2} - \frac{(1 - n_k)(1 - 2\alpha_k)^2}{(1 - q_k)^2} \right). \quad (17)$$

Evaluating these at $p_1 = 0$ and taking the expectation $\mathbb{E}_{\mathbf{n}|p=0}[n_k] = \alpha_k$, we have

$$\mathbb{E}_{\mathbf{n}|p=0} \left[\frac{\partial}{\partial p_1} \log \Pr(\mathbf{n}|p_1) \right]_{p_1=0} = 0, \quad (18)$$

$$\mathbb{E}_{\mathbf{n}|p=0} \left[\frac{\partial^2}{\partial p_1^2} \log \Pr(\mathbf{n}|p_1) \right]_{p_1=0} = \sum_k \frac{(1 - 2\alpha_k)^2}{\alpha_k(1 - \alpha_k)}. \quad (19)$$

Putting everything together in the Taylor series, we obtain

$$\mathbb{E}_{\mathbf{n}|p=0} [\lambda(\mathbf{n})] = \sum_{k=1}^N \frac{(1 - 2\alpha_k)^2}{\alpha_k(1 - \alpha_k)} p_1^2 + O(p_1^3), \quad (20)$$

where the $O(p_1^3)$ term does not scale with N w.r.t. the leading order term.

To simplify this quantity, we define the per-sample “resolution” β_k ,

$$\beta_k = \frac{(1 - 2\alpha_k)^2}{\alpha_k(1 - \alpha_k)},$$

(which, at least for small α_k , is approximately $1/\alpha_k$). The expectation value in Eq. 20 can be written concisely in terms of the average β , $\bar{\beta} = N^{-1} \sum_{k=1}^N \beta_k$, as

$$\mathbb{E}_{\mathbf{n}|p=0} [\lambda(\mathbf{n})] \sim N \bar{\beta} p_1^2.$$

Finally, we set

$$p_1 = \frac{1}{\sqrt{\bar{\beta}}} \frac{1}{\sqrt{N}}, \quad (21)$$

which ensures that $p_1 \rightarrow 0$ as $N \rightarrow \infty$ and justifies truncating the series expansion Eq. 20 above at leading order. This yields a lower bound on the minimax risk of

$$\bar{d}_{\max} \geq \bar{d}(\pi) \geq \frac{1}{2\sqrt{e\bar{\beta}}} \frac{1}{\sqrt{N}}. \quad (22)$$

It is worth noting that the risk is *not* determined by the average value of α (the per-sample noise probability), but by the average of β , which behaves roughly like $1/\alpha$. In particular, if any constant fraction of the samples are observed noiselessly, then those samples have $\beta = \infty$, and they dominate the minimax risk – $\bar{\beta} \rightarrow \infty$, and the minimax risk collapses to $O(1/N)$, as is appropriate for a noiseless coin.

Minimax risk for quantum tomography

In this section, we derive a lower bound for the minimax risk of qubit state estimation using the same framework that we used for the noisy coin. The difficulty in doing this is that a qubit’s state space (the Bloch sphere) is more complex than that of a coin – instead of a single parameter (p) there are three (x, y, z). However, the minimax risk is dominated by (i) states ρ that are very close to pure, and (ii) errors in estimating the *spectrum* of ρ (rather than errors in its eigenvectors, which contribute much less to the risk). This observation allows us to simplify the analysis greatly by choosing a bimodal prior for the qubit, supported on two states that differ only in their eigenvalues. In this circumstance, each measurement provides information equivalent (in its effect on the final estimate) to a noisy coin flip whose noisiness depends on what was measured (or, most generally, on which outcome was observed). Because we have chosen a very simple prior that is not least favorable, our analysis only guarantees a lower bound. However, it captures the dominant component of the minimax risk, and (for such a simple model) turns out to be surprisingly close to tight.

Suppose we are given N samples of a qubit state ρ . The state is drawn from a bimodal prior supported on (i) a pure state $\rho_0 = |\psi\rangle\langle\psi|$, and (ii) a slightly more mixed state $\rho_1 = (1 - p_1) |\psi\rangle\langle\psi| + p_1(\mathbb{1} - |\psi\rangle\langle\psi|)$:

$$\pi(\rho) = \frac{1}{2} (\delta(\rho - \rho_0) + \delta(\rho - \rho_1)). \quad (23)$$

We will specify $|\psi\rangle$ and $p_1 \approx 1/\sqrt{N}$ later. Each sample is measured in some basis; on the k th sample we perform the [orthogonal basis] POVM $\{|\phi_k\rangle\langle\phi_k|, \mathbb{1} - |\phi_k\rangle\langle\phi_k|\}$, and list the outcomes as a binary vector $\mathbf{n} := \{n_k\}$.

The likelihood function for a single observation is

$$\Pr(n_k|\rho_0) = |\langle\phi_k|\psi\rangle|^2 =: \alpha_k \quad (24)$$

$$\Pr(n_k|\rho_1) = (1 - 2p_1)|\langle\phi_k|\psi\rangle|^2 + p_1 = (1 - 2\alpha_k)p_1 + \alpha_k. \quad (25)$$

These are identical to the likelihoods for the noisy coin. The Bayes estimator is

$$\hat{\rho}(\mathbf{n}) = \frac{[\Pr(\mathbf{n}|\rho_0) + (1 - 2p_1)\Pr(\mathbf{n}|\rho_1)]|\psi\rangle\langle\psi| + p_1\Pr(\mathbf{n}|\rho_1)\mathbb{1}}{\Pr(\mathbf{n}|\rho_0) + \Pr(\mathbf{n}|\rho_1)}. \quad (26)$$

Now, to compute the expected risk, we observe that the Bayes estimate is always of the form $\hat{\rho} = \alpha|\psi\rangle\langle\psi| + \beta\mathbb{1}$, with

$$\alpha = \frac{[\Pr(\mathbf{n}|\rho_0) + (1 - 2p_1)\Pr(\mathbf{n}|\rho_1)]}{\Pr(\mathbf{n}|\rho_0) + \Pr(\mathbf{n}|\rho_1)}, \quad \beta = \frac{p_1\Pr(\mathbf{n}|\rho_1)}{\Pr(\mathbf{n}|\rho_0) + \Pr(\mathbf{n}|\rho_1)}. \quad (27)$$

and for any such mixture $\sigma = \alpha|\psi\rangle\langle\psi| + \beta\mathbb{1}$, the relative entropy can be computed as

$$D(|\psi\rangle\langle\psi| \parallel \sigma) = -\langle\psi| \log \sigma |\psi\rangle \quad (28)$$

$$= -\langle\psi| \log(\alpha + \beta) |\psi\rangle\langle\psi| + \log \beta (\mathbb{1} - |\psi\rangle\langle\psi|) |\psi\rangle \quad (29)$$

$$= -\log(\alpha + \beta). \quad (30)$$

Thus, in the limit of $p_1 \rightarrow 0$ and $N \rightarrow \infty$, the risk *given* that $\rho = \rho_0$ is given by

$$D(\rho_0|\hat{\rho}(\mathbf{n})) = -\log \left[\frac{\Pr(\mathbf{n}|\rho_0) + \Pr(\mathbf{n}|\rho_1)(1 - p_1)}{\Pr(\mathbf{n}|\rho_0) + \Pr(\mathbf{n}|\rho_1)} \right] \quad (31)$$

$$= -\log \left[1 - \frac{p_1\Pr(\mathbf{n}|\rho_1)}{\Pr(\mathbf{n}|\rho_0) + \Pr(\mathbf{n}|\rho_1)} \right] \quad (32)$$

$$= p_1 \frac{\Pr(\mathbf{n}|\rho_1)}{\Pr(\mathbf{n}|\rho_0) + \Pr(\mathbf{n}|\rho_1)} + O(p_1^2). \quad (33)$$

This is identical to the risk of the noisy coin.

As in Eq. 21, we choose

$$p_1 = \frac{1}{\sqrt{\bar{\beta}}} \frac{1}{\sqrt{N}}, \quad (34)$$

where $\bar{\beta}$ is defined in Equation 7 as

$$\bar{\beta} = \frac{1}{N} \sum_{k=1}^N \beta_k = \frac{1}{N} \sum_{k=1}^N \frac{(1 - 2\alpha_k)^2}{\alpha_k(1 - \alpha_k)}.$$

This yields a near-final lower bound of

$$\bar{d}_{\max} \geq r(\pi) \geq \frac{e^{-\frac{1}{2}}}{2\sqrt{\bar{\beta}}} \frac{1}{\sqrt{N}}. \quad (35)$$

To obtain a concrete lower bound on the risk, we must choose $|\psi\rangle\langle\psi|$. To ensure that the average resolution $\bar{\beta}$ is as small as possible near $|\psi\rangle$, we want α_k to be uniformly large. For the case of a qubit or a rebit, the solution is to pick the state “furthest away” from all the measurement axes, which yields $\alpha_k = \frac{1}{2} \left(1 - \frac{1}{\sqrt{D}}\right)$ where $D = 2$ for a rebit and $D = 3$ for a qubit. This yields the simple result $\bar{\beta} = 4/(D - 1)$, and therefore

$$\bar{d}_{\max} \geq r(\pi) \geq \frac{e^{-\frac{1}{2}}}{4} \frac{\sqrt{D - 1}}{\sqrt{N}}. \quad (36)$$

This argument can be extended to any discrete POVM, by choosing $|\psi\rangle$ so that it is not orthogonal to any effect of the POVM. Then α_k for each k will be lower-bounded by the minimum overlap of $|\psi\rangle$ with any effect, and $\bar{\beta}$ will be finite, and so the minimax risk will scale as $1/\sqrt{N}$.

But, by exactly the same argument, the *best* nonadaptive tomographic measurement must be unitarily symmetric. And since it must also be rank-1, it is the Haar-uniform POVM whose effects include all pure states $|\phi\rangle\langle\phi|$ with the unitarily invariant measure. The analysis given so far breaks down for this uniform POVM, because no matter what $|\psi\rangle$ we choose, the measurement has effects that diagonalize it. The effective “noise”

$$\Pr(n_k|\rho_0) = |\langle\phi_k|\psi\rangle|^2 =: \alpha_k \quad (37)$$

is distributed uniformly over $[0, 1]$. If we attempt to replace the sum

$$\sum_{k=1}^N \frac{(1 - 2\alpha_k)^2}{\alpha_k(1 - \alpha_k)} =: N\bar{\beta}, \quad (38)$$

with its average, then the integral diverges and our lower bound collapses to $\bar{d} \geq 0$.

Instead, we observe that the Haar-uniform POVM can be described as a two-step process: (1) choose a Haar uniform-basis, and (2) measure in that basis. So, a tomography experiment involving N samples can be described by a sequence $[\alpha_1, \alpha_2, \dots, \alpha_N]$, in which each α_k is drawn from the uniform distribution over $[0, 1]$. The minimax risk is the [probability-weighted] average over all such sequences. We divide them into two subsets: those in which all the $\{\alpha_k\}$ lie in the interval

$$\alpha_k \in \left[\frac{1}{2N}, 1 - \frac{1}{2N} \right]$$

and those in which at least one does not.

The probability that any given α_k lies outside the interval is exactly $1/N$, so *all* of them lie within it with probability

$$p = \left(1 - \frac{1}{N}\right)^N \geq \frac{1}{e}.$$

Conditional on all the $\{\alpha_k\}$ lying within the interval, the minimax risk can be lower bounded by integrating β over the interval, which yields

$$\bar{\beta} = 2 \log N + O(1). \quad (39)$$

This happens with probability at least $1/e$, so a lower bound on the minimax risk for the Haar-uniform POVM (as $N \rightarrow \infty$) is

$$\bar{d} \geq \frac{1}{(2e)^{3/2}} \frac{1}{\sqrt{N \log N}}. \quad (40)$$

Least favorable priors

The “Optimization Toolbox” in MATLAB 2011a, for example, contains a method `fminimax` which directly solves the optimization problem we are interested in. However, finding minimax estimators by brute force seems impossibly difficult. There are uncountably many estimators; each one is a density-matrix valued function on the set of all possible datasets. Each estimator’s performance is quantified by maximizing its risk profile $\bar{d}(\rho)$ over all density matrices ρ . Even computing the maximum risk of a single specified estimator is nontrivial; finding its minimum over the uncountable set of *all* estimators seems intractable.

Fortunately, we have some useful mathematical tools that simplify matters greatly (see, for example, [8]):

1. The minimax estimator is *also* the Bayes estimator for some measure. This fact is called *Minimax-Bayes duality*, and the measure in question is called a *least favorable prior* (LFP).
2. Relative entropy is a Bregman divergence (a.k.a. strictly proper scoring rule), and therefore the Bayes estimator for any given measure μ is Bayesian mean estimation (BME).
3. The least favorable priors for this problem are (empirically) always *discrete*, with a finite number of support points. This is not proven, but it is often the case in similar problems, and is easy to verify numerically for this problem.

Minimax-Bayes duality is enormously helpful, both as a technical tool and as an aid to problem-solving. The reasoning behind this duality is fairly straightforward:

1. Any estimator involves trade-offs in accuracy, which are quantified by its risk profile $\bar{d}(\rho)$. For example, the constant estimator $\hat{\rho}(D) = \rho_0$ is exceptionally accurate *if* the true state happens to be ρ_0 ! That is, $\bar{d}(\rho_0) = 0$. No other estimator can match its accuracy at ρ_0 . But there is a price to be paid; $\bar{d}(\rho)$ is dreadfully high for any state ρ that is far from ρ_0 .
2. Averaging $\bar{d}(\rho)$ over a measure μ quantifies these tradeoffs. In order to minimize that average, the Bayes estimator for μ must achieve fairly low expected risk in regions where μ is concentrated, but can tolerate high risk where μ is sparse.
3. If we consider the Bayes estimator for a specific measure μ_0 , its risk profile $\bar{d}(\rho)$ will typically be non-constant – so it will have at least one maximum, which we denote ρ_0 . Now suppose that we modify μ_0 (to μ') by slightly increasing the probability density around ρ_0 . The new measure μ' will have a higher Bayes risk (since ρ_0 has higher-than-average risk, and is now slightly more probable). But the Bayes estimator for μ' will be slightly different as well; it will achieve a *lower* value of $\bar{d}(\rho_0)$ because by increasing the probability of ρ_0 we have increased the value of achieving low risk at ρ_0 .
4. Iterating this process defines a flow – probability flows towards high-risk states (decreasing their expected risk) and away from low-risk states (increasing their expected risk). Every step in this flow defines a new prior (and its associated Bayes estimator) with higher Bayes risk and lower maximum risk.
5. If μ is a stationary point of this flow, then the risk profile of its Bayes estimator $\hat{\rho}_\mu(D)$ is: (i) equal to a constant C on the support of μ , and (ii) no greater than C at every point *not* in the support of μ . This estimator is necessarily minimax, because:
 - No estimator can achieve lower *average* risk on μ (by the definition of Bayes estimator),
 - So no estimator can achieve lower *maximum* risk on the support of μ (since $\hat{\rho}_\mu(D)$'s risk is constant),
 - And therefore no estimator can achieve lower maximum risk over all states (since “all states” is a superset of μ 's support).
6. Such a stationary measure can occur in one of two ways. Either $\hat{\rho}_\mu(D)$ has constant risk on *all* states, or μ is supported on a discrete set. (Because $\bar{d}(\rho)$ is analytic, it cannot be constant over a limited range, which means either it is constant everywhere or it has discrete maxima).

Numerical recipes

The above argument is the basis for the numerical algorithm we used to compute LFPs, and hence the minimax estimators. Our results were generated using an implementation of Algorithm 1, a variant of the one given by Kempthorne [19]. Although this algorithm is drastically more efficient than the brute force approach, it is still insufficient to extract the asymptotic form of the scaling for the risk of minimax estimator. In particular, it takes ~week to compute the LFP shown in Figure 4 using an implementation of Algorithm 1 in MATLAB 2011a on four 2.2GHz processors.

To remedy this, we devised an efficient Monte Carlo algorithm (Algorithm 2). The core insight is that varying only the weights of the prior renders maximizing the Bayes risk a convex optimization problem. The algorithm proceeds by randomly choosing n states according to the Hilbert-Schmidt prior. Then, the Bayes risk is maximized keeping the location of the states fixed. Both upper and lower bounds on the minimax risk can be obtained. If these are not close, then we resample near those points whose weights have not be set to zero and repeat the process.

Least favorable priors produced by Algorithm 2 are noticeably different from the (more) exact solutions obtained by Algorithm 1 (Fig. 4). However, the corresponding Bayes estimators are nearly identical, and these LFPs yield very tight upper and lower bounds on \bar{d}_{\max} (see Figure 5). We conclude that the minimax risk is very insensitive to certain variations in the prior. This explains the discrepancies in the LFPs obtained via Algorithms 1-2, and also justifies our use of the estimators and risks obtained via Algorithm 2. Using this algorithm, we were able to find good approximations to the minimax risk up to $N = 192$, but this is still insufficient to clearly show the asymptotic behavior of \bar{d}_{\max} . For that purpose, we developed the “noisy coin” model.

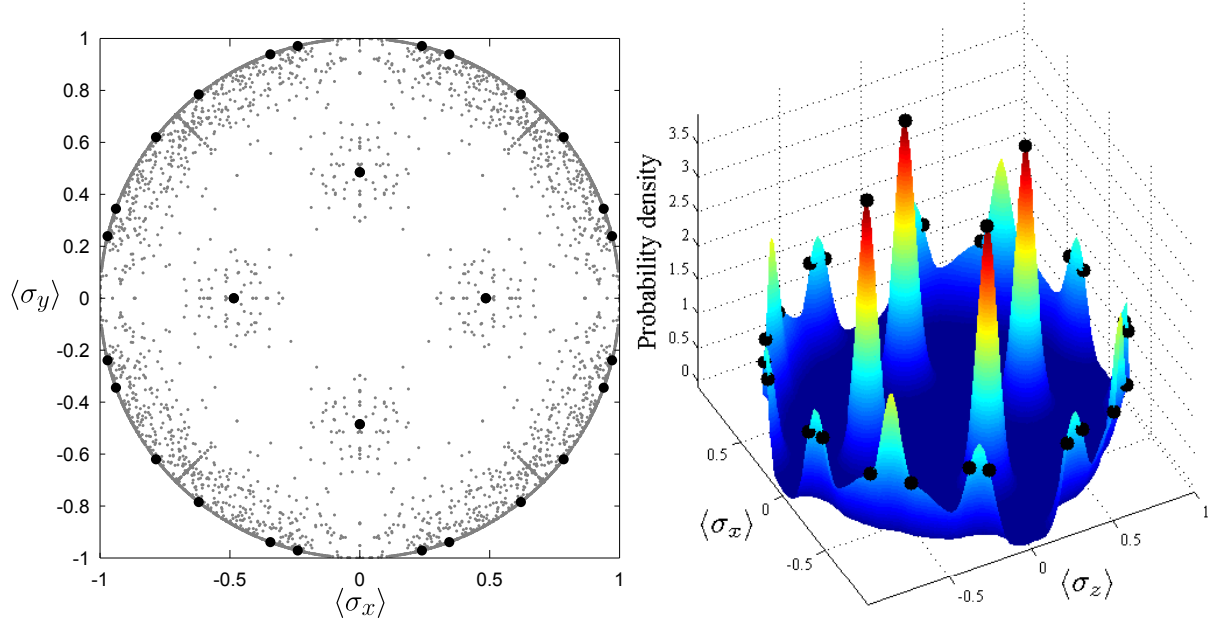


FIG. 4: Here we show the support points of numerical approximations to least favorable priors (LFPs) for $N = 16$ ($M = 8$) Pauli measurements on a *rebit* (a qubit with the constraint $\langle \sigma_y \rangle = 0$). The weights on these points are not uniform, but we shown a Gaussian kernel density estimate of them on the right. The LFP found using the highly accurate Algorithm 1 is supported on the large black dots, while the one found using the much faster Algorithm 2 is supported on the smaller gray dots. Note that in this case (and all others where we could use Algorithm 1), while the LFPs are evidently different, the resulting minimax risks are indistinguishable. We conclude that the maximum risk is insensitive to certain visible variations in the prior.

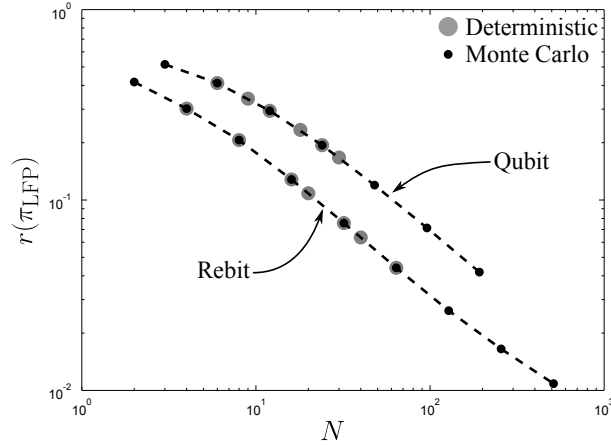


FIG. 5: Comparison of the minimax risk computed using Algorithms 1 and 2. The minimax risk of qubit and rebit tomography with N Pauli measurements ($N = 3 \dots 192$ for qubits and $N = 2 \dots 512$ for rebits) was computed by finding least favorable priors, using Algorithm 1 (large black dots) and Algorithm 2 (small gray dots). In all cases where both algorithms could be applied, results agreed to high precision.

Algorithm 1 Kempthorne (deterministic) algorithm for finding the least favorable prior [19].

Input: Number of measurements $N > 0$.
Input: Support points of initial guess prior $x_i, i \in \{1, \dots, n\}$.
Input: Probability weights of the support points $w_i, i \in \{1, \dots, n\}$ such that $\sum_i w_i = 1$.
Input: Tolerance $\text{tol} > 0$.
Input: Mixing parameter α .
Output: Least favorable prior $\{\mathbf{x}, \mathbf{w}\}$ with $m > n$ support points.
Output: Lower bound on the minimax risk av_risk .
Output: Upper bound on the minimax risk max_risk .
function DETERMINISTICLFP($N, \{\mathbf{x}, \mathbf{w}\}, \text{tol}$)
 $\text{diff} \leftarrow \text{tol} + 1$
 while $\text{diff} > \text{tol}$ **do**
 $\{\mathbf{x}, \mathbf{w}\} \leftarrow$ prior with same number of support points which maximizes the Bayes risk
 $\text{av_risk} \leftarrow$ the maximum value of the Bayes risk for the prior found above
 $\text{max_risk} \leftarrow$ global maximum of risk using the Bayes estimator of the $(\{\mathbf{x}, \mathbf{w}\})$
 $\text{diff} \leftarrow |\text{av_risk} - \text{max_risk}| / \text{av_risk}$
 if $\text{diff} > \text{tol}$ **then**
 Add a new support where the maximum risk is attained
 $w_{\text{length}(\mathbf{x})} \leftarrow \alpha$
 for each $i \leq \text{length}(\mathbf{x}) - 1, w_i \leftarrow w_i - \alpha / (\text{length}(\mathbf{x}) - 1)$
 end if
 end while
 return $\{\mathbf{x}, \mathbf{w}\}, \text{av_risk}, \text{max_risk}$
end function

Algorithm 2 Monte Carlo algorithm for finding the least favorable prior.

Input: Number of measurements $N > 0$.
Input: Number of support points $n > 0$.
Input: Tolerance on accuracy $\text{tol} > 0$.
Input: Tolerance on the weights to remove supports $\text{weight_tol} > 0$.
Input: Number of support points to add at each iteration $m > 0$ for each current support point.
Input: Variance of normal distribution to sample new points from σ .
Output: Least favorable prior $\{\mathbf{x}, \mathbf{w}\}$ with $m > n$ support points.
Output: Lower bound on the minimax risk av_risk .
Output: Upper bound on the minimax risk max_risk .
function MCLFP($N, n, \text{tol}, \text{weight_tol}$)
 $\text{diff} \leftarrow \text{tol} + 1$
 $\{\mathbf{x}, \mathbf{w}\} \leftarrow$ uniform distribution ($w_i = 1/n$) sampled according to uniform distribution over \mathbf{x}
 while $\text{diff} > \text{tol}$ **do**
 $\mathbf{w} \leftarrow$ weights which maximize the Bayes risk keeping the support points \mathbf{x} fixed
 $\text{av_risk} \leftarrow$ the maximum value of the Bayes risk for the prior found above
 $\text{max_risk} \leftarrow$ global maximum of risk using the Bayes estimator of the $(\{\mathbf{x}, \mathbf{w}\})$
 $\text{diff} \leftarrow |\text{av_risk} - \text{max_risk}| / \text{av_risk}$
 Remove all x_i such that $w_i < \text{weight_tol}$
 if $\text{diff} > \text{tol}$ **then**
 for each x_i left **do**
 Add m new support sampled randomly from $\mathcal{N}(x_i, \sigma)$
 end for
 each $w_i \leftarrow 1 / \text{length}(\mathbf{x})$
 end if
 end while
 return $\{\mathbf{x}, \mathbf{w}\}, \text{av_risk}, \text{max_risk}$
end function
